Model selection (1)

Manel Martínez-Ramón

ECE, UNM

October 2018

The model selection problem

- The Gaussian process uses kernels or covariance functions that have free parameters or *hyperparameters*.
- A proper selection of these parameters will optimize the performance of the estimator.
- A poor selection, on the other side, will result in a bad performance.
- A squared exponential plus noise can be characterized as

$$k(\mathbf{x}, \mathbf{z}) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{z})^\top \mathbf{M}(\mathbf{x}-\mathbf{z})\right) + \sigma_n^2 \delta(|\mathbf{x}-\mathbf{z}|)$$

with parameters $\boldsymbol{\theta} = \{\mathbf{M}, \sigma_f, \sigma_n\}$

The model selection problem

- The noise parameter tends to attenuate the importance of those dimensions that contain just noise, thus smoothing the solution. Indeed, the **predicted** values for the **training** data are

$$\bar{\mathbf{f}} = \mathbf{K_f}(\mathbf{K_f} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$$

where $\mathbf{K_f}$ is the kernel computed without the noise matrix [1], having the following SVD

$$\mathbf{K_f} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top = \sum_{n=1}^{N} \lambda_n \mathbf{q}_n \mathbf{q}_n^\top$$

[1] We will use later $\mathbf{K_y} = \mathbf{K_f} + \sigma_n^2\mathbf{I}$.

The model selection problem

- The inverse matrix can be expressed as

$$(\mathbf{K_f} + \sigma_n^2 \mathbf{I})^{-1} = \mathbf{Q}(\mathbf{\Lambda} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{Q}^\top = \sum_{n=1}^{N} \frac{1}{\lambda_n + \sigma_n^2} \mathbf{q}_n \mathbf{q}_n^\top$$

- Also, vector $\mathbf{y}$, of N components, can be expressed as a function of the eigenvalues, with given projection coefficients γ_n:

$$\mathbf{y} = \sum_{n=1}^{N} \gamma_n \mathbf{q}_n$$

The model selection problem

- Then, the mean vector $\bar{\mathbf{f}}$ can be expressed as

$$\bar{\mathbf{f}} = \sum_{n=1}^{N} \lambda_n \mathbf{q}_n \mathbf{q}_n^\top \sum_{n=1}^{N} \frac{1}{\lambda_n + \sigma_n^2} \mathbf{q}_n \mathbf{q}_n^\top \sum_{n=1}^{N} \gamma_n \mathbf{q}_n$$

  By virtue of the eigenvector orthonormality, the expression can be reduced as

$$\bar{\mathbf{f}} = \sum_{n=1}^{N} \frac{\lambda_n \gamma_n}{\lambda_n + \sigma_n^2} \mathbf{q}_n$$

The model selection problem

- If a dimension is important ($\lambda_n >> \sigma_n$), then

$$\frac{\lambda_n \gamma_n}{\lambda_n + \sigma_n^2} \approx \gamma_n$$

- If the dimension contains no energy corresponding to the data ($\lambda_n << \sigma_n$), in this dimension we will only find noise. Then,

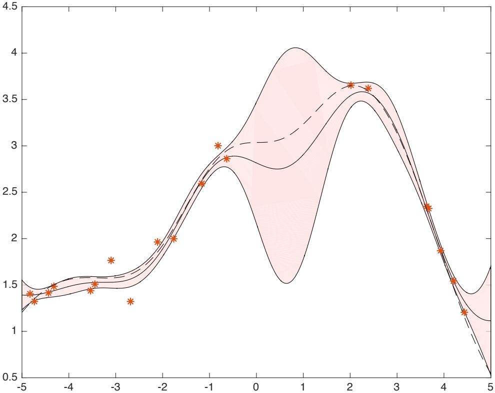
$$\frac{\lambda_n \gamma_n}{\lambda_n + \sigma_n^2} \approx \frac{\lambda_n \gamma_n}{\sigma_n^2}$$

  In that case, the noise parameter will attenuate this dimension, and it will not be used.

- This fact is a form of the optimal Wiener filtering.

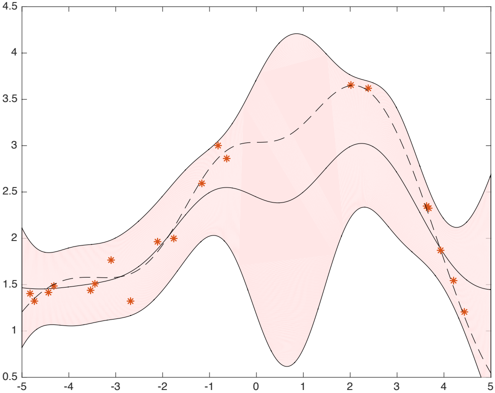
Example

- Example of regression using a squared exponential plus noise covariances where all parameters have been optimally selected. Cont line: Predicted function. Dash: Real function. 95% confidence interval.



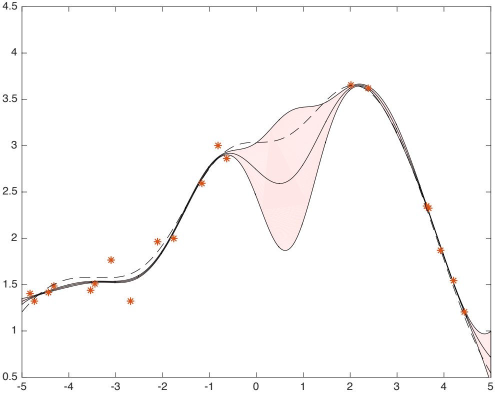
Example

- Example of regression using a squared exponential plus noise covariances where the noise parameter is 10 × the optimum one. Cont line: Predicted function. Dash: Real function. 95% confidence interval.



Example

- Example of regression using a squared exponential plus noise covariances where the noise parameter is 0.1 × the optimum one. Cont line: Predicted function. Dash: Real function. 95% confidence interval.

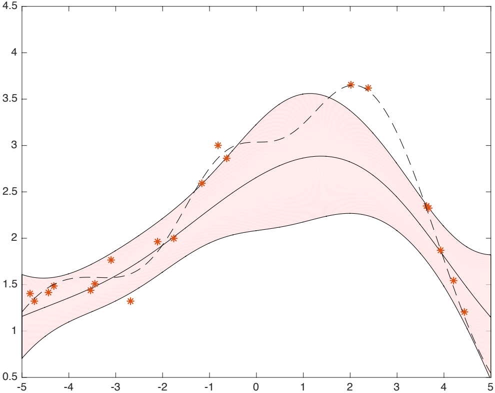


Example

- In the first example, the estimation is fair and the 95% interval contains the real function.
- In the second case, the estimation is too smooth, so the estimation error is high in some places. The confidence interval is reliable.
- In the second case, the mean seems better, because many samples have actually a low noise. Nevertheless, the confidence interval is too optimistic.
- Only if the noise parameter is known a good estimation can be achieved.

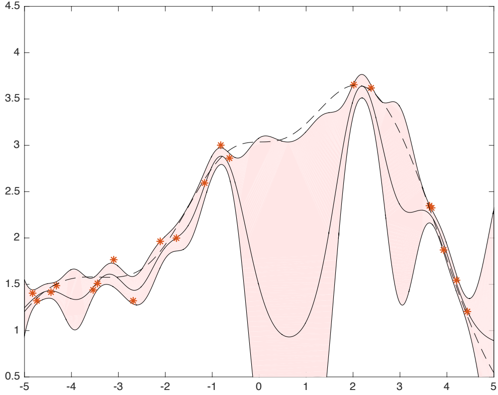
Example

- Now, the width of the square exponential is set to twice the optimal one.



Example

- And here, the width of the square exponential is set to 0.5 the optimal one.



Example

- In the first plot, we see that the solution is too smooth because the width of the kernel is too high.
- In the second plot, the solution is too complex, and it has very wide confidence intervals where the training data has low density.

The example has been uploaded as an annex to these slides. Package GPML has to be installed in Matlab.

Outcomes of this lesson

At the end of this lesson, you are required to be able to:

- Define what are the hyperparameters of a Gaussian process.
- Differentiate between kernel parameters and likelihood or noise parameter.
- Explain the model selection problem as the proper choice of the hyperparameters.
- Develop a discussion about the effects of parameter choice in all the presented examples.